

# Weak and Nested Class Memory Automata

Conrad Cotton-Barratt<sup>1,\*</sup>, Andrzej S. Murawski<sup>2,\*\*</sup>, and C.-H. Luke Ong<sup>1,\*\*\*</sup>

<sup>1</sup> Department of Computer Science, University of Oxford, UK  
 {conrad.cotton-barratt,luke.ong}@cs.ox.ac.uk

<sup>2</sup> Department of Computer Science, University of Warwick, UK  
 a.murawski@warwick.ac.uk

**Abstract** Automata over infinite alphabets have recently come to be studied extensively as potentially useful tools for solving problems in verification and database theory. One popular model of automata studied is the Class Memory Automata (CMA), for which the emptiness problem is equivalent to Petri Net Reachability. We identify a restriction – which we call weakness – of CMA, and show that they are equivalent to three existing forms of automata over data languages. Further, we show that in the deterministic case they are closed under all Boolean operations, and hence have an EXPSpace-complete equivalence problem. We also extend CMA to operate over multiple levels of nested data values, and show that while these have undecidable emptiness in general, adding the weakness constraint recovers decidability of emptiness, via reduction to coverability in well-structured transition systems. We also examine connections with existing automata over nested data.

## 1 Introduction

A data word is a word over a finite alphabet in which every position in the word also has an associated *data value*, from an infinite domain. Data languages provide a useful formalism both for problems in database theory and verification [13,16,2]. For example, data words can be used to model a system of a potentially unbounded number of concurrent processes: the data values are used as identifiers for the processes, and the data word then gives an interleaving of the actions of the processes. Having expressive, decidable logics and automata over data languages then allows properties of the modelled system to be checked.

Class memory automata (CMA) [2] are a natural form of automata over data languages. CMA can be thought of as finite state machines extended with the ability, on reading a data value, to remember what state the automaton was in when it last saw that data value. A run of a CMA is accepting if the following two conditions hold: (i) the run ends in a *globally accepting* state; and (ii) each data value read in the run was last seen in a *locally accepting* state. If using

---

\* Supported by an EPSRC Doctoral Training Grant

\*\* Supported by EPSRC (EP/J019577/1)

\*\*\* Partially supported by Merton College Research Fund

data values to distinguish semi-autonomous parts of a system, while the first condition can check the system as a whole has behaved correctly, the second of these conditions can be used to check that each part of the system independently behaved correctly. The emptiness problem for class memory automata is equivalent to Petri net reachability, and while closed under intersection, union, and concatenation, they are not closed under complementation, and do not have a decidable equivalence problem.

We earlier described how data words can be used to model concurrent systems: each process can be identified by a data value, and class memory automata can then verify properties of the system. What happens when these processes can spawn subprocesses, which themselves can spawn subprocesses, and so on? In these situations the parent-child relationship between processes becomes important, and a single layer of data values cannot capture this; instead we want a notion of *nested* data values, which themselves contain the parent-child relationship. In fact, such nested data values have applications beyond just in concurrent systems: they are prime candidates for modelling many computational situations in which names are used hierarchically. This includes higher-order computation where intermediate functional values are being created and named, and later used by referring to these names. More generally, this feature is characteristic of numerous encodings into the  $\pi$ -calculus [14].

This paper is concerned with finding useful automata models which are expressive enough to decide properties we may wish to verify, as well as having good closure and decidability properties, which make them easy to abstract our queries to. We study a restriction of class memory automata, which we find leads to improved complexity and closure results, at the expense of expressivity. We then extend class memory automata to a nested data setting, and find a decidable class of automata in this setting.

*Contributions.* In Section 3 we identify a natural restriction of Class Memory Automata, which we call *weak* Class Memory Automata, in which the local-acceptance condition of CMA is dropped. We show that these weak CMA are equivalent to: (i) Class Counting Automata, which were introduced in [11]; (ii) non-reset History Register Automata, introduced in [19]; and (iii) locally prefix-closed Data Automata, introduced in [4].

These automata have an EXPSpace-complete emptiness problem. The primary advantage of having this equivalent model as a kind of Class Memory Automaton is that there is a natural notion of determinism, and we show that Deterministic Weak CMA are closed under all Boolean operations (and hence have decidable containment and equivalence problems).

In Section 4 we extend CMA to multiple levels of “nested data”. This extension is Turing-powerful in general, but reintroducing the Weakness constraint recovers decidability. We show how these Nested Data CMA recognise the same string languages as Higher-Order Multicounter Automata, introduced in [3], and also how the weakness constraint corresponds to a natural weakness constraint on these Higher-Order Multicounter Automata. Finally, we show these automata to be equivalent to the Nested Data Automata introduced in [4].

*Related Work.* Class memory automata are equivalent to data automata (introduced in [17]), though unlike data automata, they admit a notion of determinism. Data automata (and hence class memory automata) were shown in [17] to be equiexpressive with the two-variable fragment of existential monadic second order logic over data words. Temporal logics have also been studied over data words [5], and the introduction of locally prefix-closed data automata and of nested data automata in [4] is motivated by extensions to BD-LTL, a form of LTL over multiple data values introduced in [8].

Fresh register automata [18] are a precursor to the History Register automata [19] which we examine a restriction of in this paper. Class counting automata, which we show to be equivalent to weak CMA in this paper, have been extended to be equiexpressive with CMA by adding resets and counter acceptance conditions [11,10].

We note that our restriction of class memory automata, which we call weak class memory automata, sound similar to the weak data automata introduced in [7]. However, these are two quite different restrictions, with emptiness problems of different complexities, and the two automata models should not be confused.

In the second part of this paper we examine automata over nested data values. First-order logic over nested data values has been studied in [2], where it was shown that the  $<$  predicate quickly led to undecidability, but that only having the  $+1$  predicate preserved decidability. They also examined the link between nested data and shuffle expressions. In [4] Decker et al. introduced ND-LTL, extending BD-LTL to nested data values. To show decidability of certain fragments of ND-LTL they extended data automata to run over nested data values, giving the nested data automata we examine in this paper. We note that the nested systems we introduce and study in this paper are all encodable in nested Petri nets [9].

## 2 Preliminaries

Let  $\Sigma$  be a finite alphabet, and  $\mathcal{D}$  be an infinite set of data values. A *data alphabet*,  $\mathbb{D}$ , is of the form  $\Sigma \times \mathcal{D}$ . The set of finite data words over  $\mathbb{D}$  is denoted  $\mathbb{D}^*$ . The string-projection of a word in  $\mathbb{D}^*$  is the projection to its  $\Sigma$ -values. We write this function  $str()$ , and extend it to languages over data words in the natural way. In what follows we define the automata models that will be discussed in the paper.

**Class Memory Automata and Data Automata.** Given a set  $S$ , we write  $S_\perp$  to mean  $S \cup \{\perp\}$ , where  $\perp$  is a distinguished symbol (representing a fresh data value). A *Class Memory Automaton* [2] is a tuple  $\langle Q, \Sigma, q_I, \delta, F_L, F_G \rangle$  where  $Q$  is a finite set of states,  $\Sigma$  is a finite alphabet,  $q_I \in Q$  is the initial state,  $F_G \subseteq F_L \subseteq Q$  are sets of globally- and locally-accepting sets (respectively), and  $\delta$  is the transition map  $\delta : Q \times \Sigma \times Q_\perp \rightarrow \mathcal{P}(Q)$ . The automaton is deterministic if each set in the image of the transition function is a singleton. A *class memory function* is a map  $f : \mathcal{D} \rightarrow Q_\perp$  such that  $f(d) \neq \perp$  for only finitely many  $d \in \mathcal{D}$ . We view  $f$  as a record of the history of computation: it holds the state of the automaton after the data value  $d$  was last read, where  $f(d) = \perp$  means that  $d$  is

fresh. A configuration of the automaton is a pair  $(q, f)$  where  $q \in Q$  and  $f$  is a class memory function. The initial configuration is  $(q_0, f_0)$  where  $f_0(d) = \perp$  for every  $d \in \mathcal{D}$ . Suppose  $(a, d) \in \Sigma \times \mathcal{D}$  is the input. The automaton can transition from configuration  $(q, f)$  to configuration  $(q', f')$  just if  $q' \in \delta(q, a, f(d))$  and  $f' = f[d \mapsto q']$ . A data word  $w$  is *accepted* by the automaton just if the automaton can make a sequence of transitions from the initial configuration to a configuration  $(q, f)$  where  $q \in F_G$  and  $f(d) \in F_L \cup \{\perp\}$  for every data value  $d$ .

A *Data Automaton* [17] is a pair  $(\mathcal{A}, \mathcal{B})$  where  $\mathcal{A}$  is a letter-to-letter string transducer with output alphabet  $\Gamma$ , called the Base Automaton, and  $\mathcal{B}$  is a NFA with input alphabet  $\Gamma$ , called the Class Automaton. A data word  $w = w_1 \dots w_n \in \mathbb{D}^*$  is accepted by the automaton if there is a run of  $\mathcal{A}$  on the string-projection of  $w$  (to  $\Sigma$ ) with output  $b_1 \dots b_n$  such that for each maximal set of positions  $\{x_1, \dots, x_k\} \subseteq \{1, \dots, n\}$  such that  $w_{x_1}, \dots, w_{x_k}$  share the same data value, the word  $b_{x_1} \dots b_{x_k}$  is accepted by  $\mathcal{B}$ .

CMA and DA are expressively equivalent, with PTIME translation [2]. The emptiness problem for these automata is decidable, and equivalent to Petri Net Reachability [17]. The class of languages recognised by CMA is closed under intersection, union, and concatenation. It is not closed under complementation or Kleene star. Of the above, the class of languages recognised by deterministic CMA is closed only under intersection.

**Locally Prefix-Closed Data Automata.** A Data Automaton  $\mathcal{D} = (\mathcal{A}, \mathcal{B})$  is locally prefix-closed (pDA) [4] if all states in  $\mathcal{B}$  are final. The emptiness problem for pDA is EXPSpace-complete [4].

**Class Counting Automata.** A *bag* over  $\mathcal{D}$  is a function  $h : \mathcal{D} \rightarrow \mathbb{N}$  such that  $h(d) = 0$  for all but finitely many  $d \in \mathcal{D}$ . Let  $C = \{=, \neq, <, >\} \times \mathbb{N}$ , which we call the set of constraints. If  $c = (\text{op}, e) \in C$  and  $n \in \mathbb{N}$  we write  $n \models c$  iff  $n \text{ op } e$ . A *Class Counting Automaton* (CCA) [11] is a tuple  $\langle Q, \Sigma, \Delta, q_0, F \rangle$  where  $Q$  is a finite set of states,  $\Sigma$  is a finite alphabet,  $q_0$  is the initial state,  $F \subseteq Q$  is the set of accepting states, and  $\Delta$ , the transition relation, is a finite subset of  $Q \times \Sigma \times C \times \{\uparrow^+, \downarrow\} \times \mathbb{N} \times Q$ . A configuration of a CCA,  $\mathcal{C} = \langle Q, \Sigma, \Delta, q_0, F \rangle$ , is a pair  $(q, h)$  where  $q \in Q$  and  $h$  is a bag. The initial configuration is  $(q_0, h_0)$  where  $h_0$  is the zero function. Given a data word  $w = (a_1, d_1)(a_2, d_2) \dots (a_n, d_n)$  a run of  $w$  on  $\mathcal{C}$  is a sequence of configurations  $(q_0, h_0)(q_1, h_1) \dots (q_n, h_n)$  such that for all  $0 \leq i < n$  there is a transition  $(q, a, c, \pi, m, q')$  where  $q = q_i$ ,  $q' = q_{i+1}$ ,  $a = a_{i+1}$ ,  $h_i(d_{i+1}) \models c$ , and

$$h_{i+1} = \begin{cases} h_i[d_{i+1} \mapsto h_i(d_{i+1}) + m] & \text{if } \pi = \uparrow^+ \\ h_i[d_{i+1} \mapsto m] & \text{if } \pi = \downarrow \end{cases}$$

The run is accepting if  $q_n \in F$ . The emptiness problem for Class Counting Automata was shown to be EXPSpace-complete in [11].

**Non-Reset History Register Automata.**<sup>1</sup> For a positive integer  $k$  write  $[k]$  for the set  $\{1, 2, \dots, k\}$ . Fixing a positive integer  $m$ , define the set of labels

<sup>1</sup> We provide a simplified definition to that provided in [19], since we do not need to consider full History Register Automata. In particular, due to Proposition 22 in [19], we need only consider histories, and not registers.

$\text{Lab} = \mathcal{P}([m])^2$ . A non-reset History Register Automaton (nrHRA) of type  $m$  with initially empty assignment is a tuple  $\mathcal{A} = \langle Q, \Sigma, \delta, q_0, F \rangle$  where  $q_0 \in Q$  is the initial state,  $F \subseteq Q$  is the set of final states, and  $\delta \subseteq Q \times \Sigma \times \text{Lab} \times Q$ . A configuration of  $\mathcal{A}$  is a pair  $(q, H)$  where  $q \in Q$  and  $H : [m] \rightarrow \mathcal{P}_{fn}(\mathcal{D})$  where  $\mathcal{P}_{fn}(\mathcal{D})$  is the set of finite subsets of  $\mathcal{D}$ . We call  $H$  an *assignment*, and for  $d \in \mathcal{D}$  we write  $H^{-1}(d)$  for the set  $\{i \in [m] : d \in H(i)\}$ . The initial configuration is  $(q_0, H_0)$ , where  $H_0$  assigns every integer in  $[m]$  to the empty set. When the automaton is in configuration  $(q, H)$ , on reading input  $\begin{pmatrix} a \\ d \end{pmatrix}$  it can transition to configuration  $(q', H')$  providing there exists  $X \subseteq [m]$  such that  $(q, a, (H^{-1}(d), X), d) \in \delta$  and  $H'$  is obtained by removing  $d$  from  $H(i)$  for each  $i$  then adding  $d$  to each  $H(i)$  such that  $i \in X$ . A run is defined in the usual way, and a run is accepting if it ends in a configuration  $(q, H)$  where  $q \in F$ .

**Higher-Order Multicounter Automata.** A multiset over a set  $A$  is a function  $m : A \rightarrow \mathbb{N}$ . A level-1 multiset over  $A$  is a finite multiset over  $A$ . A level- $(k+1)$  multiset over  $A$  is a finite multiset of level- $k$  multisets over  $A$ . We can visualise this with nested set notation: e.g.  $\{\{a, a\}, \{\}, \{\}\}$  represents the level-2 multiset containing one level-1 multiset containing two copies of  $a$ , and two empty level-1 multisets. A multiset is *hereditarily empty* if, written in nested set notation, it contains no symbols from  $A$ .

Higher-Order Multicounter Automata (HOMCA) were introduced in [3], and their emptiness problem was shown to be Turing-complete at level-2 and above. A level- $k$  multicounter automaton is a tuple  $\langle Q, \Sigma, A, \Delta, q_0, F \rangle$  where  $Q$  is a finite set of states,  $\Sigma$  is the input alphabet,  $A$  is the multiset alphabet,  $q_0$  is the initial state, and  $F$  is the set of final states. A configuration is a tuple  $(q, m_1, m_2, \dots, m_k)$  where  $q \in Q$  and each  $m_i$  is either undefined ( $\perp$ ) or a level- $i$  multiset over  $A$ . The initial configuration is  $(q_0, \perp, \dots, \perp)$ .  $\Delta$  is the transition relation, and is a subset of  $Q \times \Sigma \times \text{ops} \times Q$  where *ops* is the set of possible counter operations. These operations, and meanings, are as follows: (i) *new<sub>i</sub>* ( $i \leq k$ ) turns  $m_i$  from  $\perp$  into the empty level- $i$  multiset; (ii) *inc<sub>a</sub>* ( $a \in A$ ) adds  $a$  to  $m_1$ ; (iii) *dec<sub>a</sub>* ( $a \in A$ ) removes  $a$  from  $m_1$ ; (iv) *store<sub>i</sub>* ( $i < k$ ) adds  $m_i$  to  $m_{i+1}$  and sets  $m_i$  to  $\perp$ ; (v) *load<sub>i</sub>* ( $i < k$ ) non-deterministically removes an  $m$  from  $m_{i+1}$  and turns  $m_i$  from  $\perp$  to  $m$ . This can happen only when  $m_1 \dots m_i$  are all  $\perp$ . The automaton reads the input word from left to right, updating  $m_1 \dots m_k$  as determined by the transitions. A word is accepted by the automaton just if there is a run of the word such that the automaton ends up in configuration  $(q, m_1, \dots, m_k)$  where  $q \in F$  and each  $m_i$  is hereditarily empty.

### 3 Weak Class Memory Automata

In this section we introduce a restriction of class memory automata, *weak* class memory automata (WCMA), and discuss the improved closure and complexity properties. We also show that WCMA correspond to a natural restriction of data automata, locally-prefix closed data automata, as well as two other independent automata models, class counting automata and non-reset history register automata.

**Definition 1.** A class memory automaton  $\langle Q, \Sigma, \Delta, q_0, F_L, F_G \rangle$  is weak if all states are locally accepting (i.e.  $F_L = Q$ ).

When defining a weak CMA (WCMA) we may omit the set of locally accepting states, and just give one set of final states,  $F$ .

The emptiness problem for class memory automata is reducible (in fact, equivalent) to emptiness of multicounter automata (MCA) [17,2]. This reduction works by using counters to store the number of data values last seen in each state. The local-acceptance condition is checked by the zero-test of each counter at the end of a run of an MCA. In the weak CMA case, this check is no longer necessary, and so emptiness is reducible to emptiness of weak MCA. Just as MCA emptiness is equivalent to Petri net reachability, weak MCA emptiness is equivalent to Petri net coverability.

*Example 2.* We give an example showing how a very simple Petri net reachability query can be reduced to an emptiness of CMA problem, and the small change required to reduce coverability queries to emptiness of WCMA. The idea is to encode tokens in the Petri net using data values: the location of the token is stored by the class memory function's memory for the data value. Transitions in the Petri net will be simulated by sequences of transitions in the automaton, which change class memory function appropriately. Consider the Petri net shown in Figure 1, with initial marking on the left and target marking on the right.



**Figure 1.** An example Petri net with initial and target markings.

We give the automaton which models this reachability query in Figure 2. The first transitions from the initial state just set up the initial marking. As there is only one token in the initial marking, this just involves reading one fresh data value: this is the transition from  $s_0$  to  $s_1$  below. Once the initial marking has been set up (reaching  $s_2$  below), the automaton can simulate the transitions firing any number of times. Each loop from  $s_2$  back to itself represents one transition in the Petri net firing: the loop above represents  $t_1$  firing, and the loop below represents  $t_2$  firing. For  $t_1$  to fire, no preconditions must be met, and a new data value can be read in state  $s_3$ , thus data values last seen in either of states  $s_1$  and  $s_3$  represent tokens in  $p_1$ . For  $t_2$  to fire, a token must be removed from  $p_1$ , since tokens in  $p_1$  are represented by tokens in either  $s_1$  or  $s_3$ , the first transition in this loop – to  $s_4$  – involves reading a data value last seen in one of these states. Thus data values seen in  $s_4$  represent removed tokens, which we do not use again. Then a new token is placed in  $p_2$  by reading a fresh data value in  $s_5$ . Once back in  $s_2$  these loops can be taken more, or the fact that a marking covering the target marking has been reached can be checked by reading two data values last seen in  $s_5$  to reach a final state. The only globally accepting state is  $s_7$ , and all states except those which are used to represent tokens – i.e. all except  $s_1$ ,  $s_3$ , and  $s_5$  – are locally accepting. The local acceptance condition thus checks that no other tokens remain in the simulated Petri net.

If we were interested in a coverability query, the same automaton, but without the local acceptance condition, would obviously suffice. Thus emptiness of WCMA is equivalent to Petri net coverability, which is EXSPACE-complete.

We now give the main observation of this section: that weak CMA are equivalent to three independent existing automata models.

**Theorem 3.** *Weak CMA, locally prefix<sup>1</sup>-closed DA, class counting automata, and non-reset history register automata are all PTIME-equivalent.*

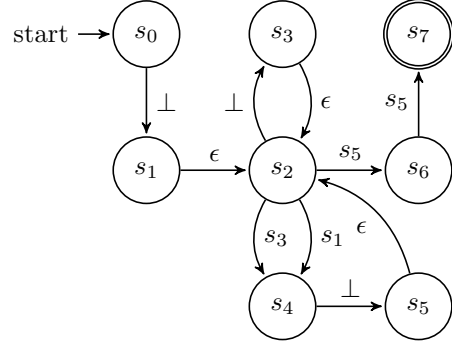
*Proof.* That Weak CMA and pDA are equivalent is a simple alteration of the proof of equivalence of CMA and DA provided in [2].

Recall that CCA use a “bag”, which essentially gives a counter for each data value. Weak CMA can easily be simulated by CCA by identifying each state with a natural number; then the bag can easily simulate the class memory function, by setting the data value’s counter to the appropriate number when it is read. To simulate a CCA with a WCMA, we first observe that for any CCA, since counter values can only be incremented or reset, there is a natural number,  $N$ , above which different counter values are indistinguishable to the automaton. Thus we need only worry about a finite set of values. This means the value for the counter of each data value can be stored in the automaton state, and thereby the class memory function. A full proof is provided in Appendix A.

In [19] the authors already show that nrHRA can be simulated by CMA. Their construction does not make use of the local-acceptance condition, so the fact that nrHRA can be simulated by WCMA is immediate. In order to simulate a given WCMA with state set  $[m]$ , one can take a nrHRA of type  $m$ , with place  $i$  storing the data values last seen in state  $i$ .

Data automata, and hence pDA, unlike CMA and WCMA, do not have a natural notion of determinism, nor a natural restriction corresponding to deterministic CMA or WCMA. What about for CCA? We define CCA to be deterministic if for each state  $q$  and input letter  $a$ , the transitions  $(q, a, c, \dots) \in \Delta$  are such that the  $c$ ’s partition  $\mathbb{N}$ . The translations provided in Appendix A also show that deterministic WCMA and deterministic CCA are equivalent. We can ask the same question of non-reset HRA. We find that the natural notion of determinism here is: for each  $q \in Q$ ,  $a \in \Sigma$ , and  $X \subseteq [m]$  there is precisely one  $Y \subseteq [m]$  and  $q' \in Q$  such that  $(q, a, (X, Y), q') \in \delta$ . Similarly, the translations discussed above show deterministic WCMA to be equivalent to deterministic nrHRA.

It follows from the results for CCA in [11] that Weak CMA, like normal CMA, are closed under intersection and union, though these closures can easily



**Figure 2.** A class memory automaton simulating the Petri net query shown in Figure

be shown directly using product constructions. In fact, Deterministic Weak CMA have even nicer closure properties:

**Proposition 4.** *Deterministic Weak CMA are closed under all Boolean operations.*

*Proof.* Closure under intersection and union can be shown by product constructions. For complementation one can use the same method as for DFA: complementing the final states.

**Corollary 5.** *The containment and equivalence problems for Deterministic Weak CMA are EXPSpace-complete.*

## 4 Nested Data Class Memory Automata

In Section 1 we discussed how data values fail to provide a good model for modelling computations in which names are used hierarchically, such as a system of concurrent processes which can spawn subprocesses. Motivated by these applications, in this section we introduce a notion of nested data values in which the data set has a forest-structure. This is a stylistically different presentation to earlier work on nested data in that [2,4] require that each position in the words considered have a data value in each of a fixed number of levels. By giving the data set a forest-structure, we can explicitly handle variable levels of nesting within a word. However, we note that there is a natural translation between the two presentations.

**Definition 6.** *A rooted tree (henceforth, just tree) is a simple directed graph  $\langle D, \text{pred} \rangle$ , where  $\text{pred} : D \rightarrow D$  is the predecessor map defined on every node of the tree except the root, such that every node has a unique path to the root. A node  $n$  of a tree has level  $l$  just if  $\text{pred}^{l-1}(n)$  is the root (thus the root has level 1). A tree has bounded level just if there exists a least  $l \geq 1$  such that every node has level no more than  $l$ ; we say that such a tree has level  $l$ .*

*We define a nested dataset  $\langle \mathcal{D}, \text{pred} \rangle$  to be a forest of infinitely many trees of level  $l$  which is full in the sense that for each data value  $d$  of level less than  $l$ ,  $d$  has infinitely many children.*

We now extend CMA to nested data by allowing the nested data class memory automaton (NDCMA), on reading a data value  $d$ , to access the class memory function's memory of not only  $d$ , but each ancestor of  $d$  in the nested data set. Once a transition has been made, the class memory function updates the remembered state not only of  $d$ , but also of each of its ancestors. Formally:

**Definition 7.** *Fix a nested data set of level  $l$ . A Nested Data CMA of level  $l$  is a tuple  $\langle Q, \Sigma, \delta, q_0, F_L, F_G \rangle$  where  $Q$  is a finite set of states,  $q_0 \in Q$  is the initial state,  $F_G \subseteq F_L \subseteq Q$  are sets of globally and locally accepting states respectively, and  $\delta$  is the transition map.  $\delta$  is given by a union  $\delta = \bigcup_{1 \leq i \leq l} \delta_i$  where each  $\delta_i$  is a function:*

$$\delta_i : Q \times \Sigma \times (\{i\} \times (Q_\perp)^i) \rightarrow \mathcal{P}(Q)$$



The automaton is deterministic if each set in the image of  $\delta$  is a singleton; and is weak if  $F_L = Q$ . A configuration is a pair  $(q, f)$  where  $q \in Q$ , and  $f : \mathcal{D} \rightarrow Q_\perp$  is a class memory function (i.e.  $f(d) = \perp$  for all but finitely many  $d \in \mathcal{D}$ ). The initial configuration is  $(q_0, f_0)$  where  $f_0$  is the class memory function mapping every data value to  $\perp$ . A configuration  $(q, f)$  is final if  $q \in F_G$  and  $f(d) \in F_L \cup \{\perp\}$  for all  $d \in \mathcal{D}$ . The automaton can transition from configuration  $(q, f)$  to configuration  $(q', f')$  on reading input  $\binom{a}{d}$  just if  $d$  is a level- $i$  data value,  $q' \in \delta(q, a, (i, f(\text{pred}^{i-1}(d)), \dots, f(\text{pred}(d)), f(d)))$ , and  $f' = f[d \mapsto q, \text{pred}(d) \mapsto q, \dots, \text{pred}^{i-1}(d) \mapsto q]$ . A run  $(q_0, f_0), (q_1, f_1), \dots, (q_n, f_n)$  is accepting if the configuration  $(q_n, f_n)$  is final.  $w \in L(\mathcal{A})$  if there is an accepting run of  $\mathcal{A}$  on  $w$ .

It is clear that level-1 NDCMA are equivalent to normal CMA. We know that emptiness of class memory automata is equivalent to reachability in Petri nets; it is natural to ask whether there is any analogous correspondence – to some kind of high-level Petri net – once nested data is used.

*Example 8.* In Example 2, we showed how CMA (resp. WCMA) can encode Petri net reachability (resp. coverability). A similar technique allows reachability (resp. coverability) of Petri nets with reset arcs to be reduced to emptiness of NDCMA (resp. weak NDCMA). The key idea is to have, for each place in the net, a level-1 data value – essentially as a “bag” holding the tokens for that place. Nested under the level-1 data value, level-2 data values are used to represent tokens just as before. When a reset arc is fired, the corresponding level-1 data value is moved to a “dead” state – from where it and the data values nested under it are not moved again – and a fresh level-1 data value is then used to hold subsequently added tokens to that place.

**Theorem 9.** *The emptiness problem for NDCMA is undecidable. Emptiness of Weak NDCMA is decidable, but Ackermann-hard.*

*Proof.* This result follows from Theorem 13 together with results in [4], though we also provide a direct proof.

We show decidability by reduction to a well-structured transition system [6] constructed as follows: a class memory function on a nested data set can be viewed as a labelling of the data set by labels from the set of states. Since we only care about the shape of the class memory function (i.e. up to automorphisms of the nested data set), we can remove the nodes labelled by  $\perp$ , and view a class memory function as a finite set of labelled trees. The set of finite forests of finite trees of bounded depth with the order given by  $F \leq F'$  iff there is a forest homomorphism from  $F$  to  $F'$  (where a forest is the natural extension of tree homomorphisms to forests) is a well-quasi-order [12], which provides the basis for the well-structured transition system. A full proof is given in Appendix B.

Undecidability for NDCMA and Ackermann-hardness for Weak NDCMA follow from the ideas in Example 8: the reachability (resp. coverability) problem for Petri nets with reset arcs is encodable in NDCMA (resp. Weak NDCMA), and this is undecidable [1] (resp. Ackermann-hard [15]).

Weak Nested Data CMA have similar closure properties to Weak CMA.

**Proposition 10.** (i) *Weak NDCMA are closed under intersection and union.*  
(ii) *Deterministic Weak NDCMA are closed under all Boolean operations.*

*Proof.* Again these can be shown by the same techniques as for DFA.

**Corollary 11.** *The containment and equivalence problems for Deterministic Weak NDCMA are decidable.*

#### 4.1 Link with Nested Data Automata

In [4] Decker et al. also examined “Nested Data Automata” (NDA), and showed the locally prefix-closed NDA (pNDA) to have decidable emptiness (via reduction to well-structured transition systems). In fact, these NDA precisely correspond to NDCMA, and again being locally prefix-closed corresponds to weakness. In this section we briefly outline this connection.

**Nested Data Automata.** A  $k$ -nested data automaton is a tuple  $(\mathcal{A}, \mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_k)$  where  $(\mathcal{A}, \mathcal{B}_i)$  is a data automaton for each  $i$ . Such automata run on words over the alphabet  $\Sigma \times \mathcal{D}^k$ , where  $\mathcal{D}$  is a (normal, unstructured) dataset. As for normal data automata, the transducer,  $\mathcal{A}$ , runs on the string projection of the word, giving output  $w$ . Then for each  $i$  the class automaton  $\mathcal{B}_i$  runs on each subsequence of  $w$  corresponding to the positions which agree on the first  $i$  data values. The NDA is locally prefix-closed if each  $(\mathcal{A}, \mathcal{B}_i)$  is.

Since these NDA are defined on a slightly different presentation of nested data, we provide the following presentation of NDCMA over multiple levels of data.

**Definition 12.** A Nested Data CMA of level  $k$  over the alphabet  $\Sigma \times \mathcal{D}^k$  is a tuple  $\langle Q, \Sigma, \delta, q_0, F_L, F_G \rangle$  where  $Q$  is a finite set of states,  $q_0 \in Q$ ,  $F_G \subseteq F_L \subseteq Q$ , and  $\delta : Q \times \Sigma \times (Q_\perp)^k \rightarrow \mathcal{P}(Q)$  is the transition map.

A configuration is a tuple  $(q, f_1, f_2, \dots, f_k)$ , where each  $f_i : \mathcal{D}^i \rightarrow Q_\perp$  maps an  $i$ -tuple of data values to a state in the automaton (or  $\perp$ ). The initial configuration is  $(q_0, f_1^0, \dots, f_k^0)$  where  $f_i^0$  maps every tuple in the domain to  $\perp$ . A configuration  $(q, f_1, \dots, f_k)$  is final if each  $f_i$  maps into  $F_L \cup \{\perp\}$ . The automaton can transition from configuration  $(q, f_1, \dots, f_k)$  to configuration  $(q', f'_1, \dots, f'_k)$  on reading input  $(a, d_1, \dots, d_k)$  just if  $q' \in \delta(q, a, (f_1(d_1), f_2(d_1, d_2), \dots, f_k(d_1, \dots, d_k)))$ , and each  $f'_i = f_i[(d_1, \dots, d_i) \mapsto q']$ .

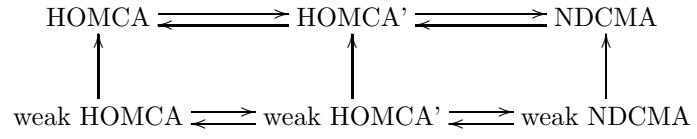
Using ideas from the proof of equivalence between CMA and DA in [2], we can show the following result:

**Theorem 13.** *NDCMA (resp. weak NDCMA) and NDA (resp. pNDA) are expressively equivalent, with effective translations.*

## 4.2 Link with Higher-Order Multicounter Automata

In [3] the authors examined a link between nested data values and shuffle expressions. In doing so, they introduced higher-order multicounter automata (HOMCA). While not explicitly over nested data values, they are closely related to the ideas involved, and in fact we show that, just as multicounter automata and CMA are equivalent, there is a natural translation between HOMCA and the NDCMA we have introduced. Further, just as the equivalence between MCA and CMA descends to one between weak NDCMA and weak CMA, we find an equivalence between weak NDCMA and “weak” HOMCA in which the corresponding acceptance condition – hereditary emptiness – is dropped. To show this, we introduce HOMCA’, which add restrictions to the  $store_i$  and  $new_i$  counter operations analogous to the restriction for the  $load_i$  operation.

We show that these HOMCA’ are equivalent to HOMCA, and that HOMCA’ are equivalent to NDCMA, with both of these equivalences descending to the weak versions. These equivalences are summarised in Figure 3.



**Figure 3.** A diagram showing translations between HOMCA, HOMCA’, NDCMA, and their weak counterparts.

**Definition 14.** We define weak HOMCA to be just as HOMCA, but without the hereditary-emptiness condition on acceptance, i.e. a run is accepting just if it ends in a final state.

**Definition 15.** We define HOMCA’ to be the same as HOMCA, except for the following changes to the counter operations: (i)  $store_i$  operations are only enabled when  $m_1 = m_2 = \dots = m_{i-1} = \perp$ ; and (ii)  $new_i$  operations are only enabled when  $m_k \neq \perp, m_{k-1} \neq \perp, \dots, m_{i+1} \neq \perp$  and  $m_{i-1} = m_{i-2} = \dots = m_1 = \perp$ .

As for HOMCA, we define weak HOMCA’ to be HOMCA’ without the hereditary-emptiness condition.

This means that each reachable configuration  $(q, m_1, \dots, m_k)$  is such that there is a unique  $0 \leq i \leq k$  such that for all  $j \leq i$ ,  $m_j = \perp$  and each  $l > i$ ,  $m_l \neq \perp$ .

**Theorem 16.** HOMCA (resp. weak HOMCA) and HOMCA’ (resp. weak HOMCA’) are expressively equivalent, with effective translations between them.

*Proof.* This requires simulating the HOMCA operations  $store_i$  and  $new_i$  in HOMCA’: which can be difficult if, for instance, the HOMCA is carrying out a  $store_i$  operation when it has a current level- $(i-1)$  multiset in memory. The trick is to move the level- $(i-1)$  multiset across to be nested under a new level- $i$  multiset, and this can be done one element at a time in a “folding-and-unfolding” method. The hereditary emptiness condition checks that each of these movements was completed, i.e. no element was left unmoved. In the weak case some

elements not being moved could not change an accepting run to a non-accepting run, so the fallibility of the moving method does not matter. We provide a more detailed proof sketch in Appendix C.

We now have the main result of this section:

**Theorem 17.** *For every (weak) level- $k$  NDCMA,  $\mathcal{A}$ , there is a (weak) level- $k$  HOMCA',  $\mathcal{A}'$ , such that  $\mathcal{L}(\mathcal{A}') = \text{str}(\mathcal{L}(\mathcal{A}))$ , and vice-versa.*

*Proof.* This proof rests on the strong similarity between the nesting of data values, and the nesting of level- $i$  multisets in level- $(i + 1)$  multisets.

For a NDCMA to simulate a HOMCA', we use level- $k$  data values to represent instances of the multiset letters, level- $(k - 1)$  data values to represent level-1 multisets, and so on, up to level-1 data values representing level- $(k - 1)$  multisets. Since each run of a HOMCA' can have at most one level- $k$  multiset, this does not need to be encoded in data values.

Conversely, when simulating a NDCMA with a HOMCA', a level- $k$  data value is represented by an instance of an appropriate multiset letter. The letter contains the information on which state the data value was last seen in. Level- $(k - 1)$  data values are represented by level-1 multisets, which also include a multiset letter storing the state that data value was last seen in. We provide a full proof of the theorem in Appendix D.

## 5 Conclusion

We showed that by dropping local acceptance conditions in Class Memory Automata one obtains a robust class of languages that has already appeared in the literature under various guises. Furthermore, its deterministic restriction is closed under all Boolean operations, which implies decidability (in fact, EXPSPACE-completeness) of the containment and equivalence problems. We recall that both were undecidable for general class memory automata.

We introduce a new notion of nesting for data languages, based on tree-structured datasets. This notion does not commit all letters to be at the same level of nesting and appears promising from the point of view of modelling scenarios with hierarchical name structure, such as concurrent or higher-order computation. We extend Class Memory Automata to operate over these nested datasets, and show that without the local acceptance condition, these have a decidable emptiness problem, and in the deterministic case are closed under all Boolean operations.

In future work, we would like to understand better whether there is a natural fragment of the  $\pi$ -calculus that corresponds to the new classes of automata. On the logical side, an interesting outstanding question is to characterize languages accepted by our classes of automata with suitable logics.

## References

1. Toshiro Araki and Tadao Kasami. Some decision problems related to the reachability problem for Petri nets. *Theor. Comput. Sci.*, 3(1):85–104, 1976.
2. Henrik Björklund and Thomas Schwentick. On notions of regularity for data languages. *Theor. Comput. Sci.*, 411(4-5):702–715, 2010.
3. Henrik Björklund and Mikołaj Bojańczyk. Shuffle expressions and words with nested data. In *Proceedings of MFCS*, volume 4708 of *LNCS*, pages 750–761, 2007.
4. N. Decker, P. Habermehl, M. Leucker, and D. Thoma. Ordered navigation on multi-attributed data words. In *CONCUR 2014*, pages 497–511. Springer, 2014.
5. Stéphane Demri and Ranko Lazić. LTL with the freeze quantifier and register automata. *ACM Trans. Comput. Log.*, 10(3), 2009.
6. Alain Finkel and Ph. Schnoebelen. Well-structured transition systems everywhere! *Theor. Comput. Sci.*, 256(1-2):63–92, 2001.
7. A. Kara, T. Schwentick, and T. Zeume. Feasible automata for two-variable logic with successor on data words. In *LATA*, volume 7183 of *LNCS*, pages 351–362. Springer, 2012.
8. A. Kara, T. Schwentick, and T. Zeume. Temporal logics on words with multiple data values. In *FSTTCS*, volume 8 of *LIPIcs*, pages 481–492. Schloss Dagstuhl, 2010.
9. Irina A. Lomazova and Ph. Schnoebelen. Some decidability results for nested petri nets. In *Ershov Memorial Conference*, volume 1755 of *LNCS*, pages 208–220. Springer, 1999.
10. Amaldev Manuel. *Counter automata and classical logics for data words*. PhD thesis, Institute of Mathematical Sciences, Chennai, 2011.
11. Amaldev Manuel and Ramaswamy Ramanujam. Counting multiplicity over infinite alphabets. In *RP*, volume 5797 of *LNCS*, pages 141–153. Springer, 2009.
12. Roland Meyer. On boundedness in depth in the pi-calculus. In *IFIP TCS*, volume 273 of *IFIP*, pages 477–489. Springer, 2008.
13. F. Neven, T. Schwentick, and V. Vianu. Finite state machines for strings over infinite alphabets. *ACM Trans. Comput. Log.*, 5(3):403–435, 2004.
14. D. Sangiorgi. *Expressing Mobility in Process Algebras: First-Order and Higher-Order Paradigms*. PhD thesis, University of Edinburgh, 1992.
15. Ph. Schnoebelen. Verifying lossy channel systems has nonprimitive recursive complexity. *Inf. Process. Lett.*, 83(5):251–261, 2002.
16. Luc Segoufin. Automata and logics for words and trees over an infinite alphabet. In *CSL*, volume 4207 of *LNCS*, pages 41–57. Springer, 2006.
17. M. Bojańczyk, A. Muscholl, T. Schwentick, L. Segoufin, and C. David. Two-variable logic on words with data. In *LICS*, pages 7–16. IEEE Computer Society, 2006.
18. Nikos Tzevelekos. Fresh-register automata. In Thomas Ball and Mooly Sagiv, editors, *POPL*, pages 295–306. ACM, 2011.
19. Nikos Tzevelekos and Radu Grigore. History-register automata. In *FoSSaCS*, volume 7794 of *Lecture Notes in Computer Science*, pages 17–33. Springer, 2013.

## A Proof that WCMA and CCA are equivalent

**Proposition 18.** *Given a WCMA a CCA recognising the same language can be constructed in PTIME.*

*Proof.* Let  $\mathcal{A} = \langle Q, \Sigma, \Delta, q_0, F \rangle$  be a WCMA. WLOG suppose  $Q = \{q_1, \dots, q_n\}$ . Consider the CCA  $\mathcal{A}' = \langle Q, \Sigma, \Delta', q_0, F \rangle$  where  $\Delta'$  is given by:

$$\begin{aligned} \Delta' = & \{(q_i, a, (=, k), \downarrow, j, q_j) \mid k \geq 1 \text{ and } (q_i, a, q_k, q_j) \in \Delta\} \\ & \cup \{(q_i, a, (=, 0), \downarrow, j, q_j) \mid (q_i, a, \perp, q_j) \in \Delta\} \end{aligned}$$

The bag is used to simulate the class memory function by representing a state by a natural number. That these automata recognise the same language is a straightforward induction.

**Proposition 19.** *Given a CCA a WCMA recognising the same language can be constructed in PTIME.*

*Proof.* The idea of this construction is to use the class memory function to mimic the counters, which can be done by representing the counter value in the states. This is possible since only a finite number of counter values can behave differently: if the greatest number used in the constraints of the transition function of a CCA is  $n_0$  then all the integers above  $n_0$  are equivalent for the purposes of which transitions apply. Further, since the counters can only be incremented or reset, there is no need to track how much greater than  $n_0$  such a counter is.

Let  $\mathcal{A} = \langle Q, \Sigma, \Delta, q_0, F \rangle$  be a CCA. Since  $\Delta$  is finite, there is a greatest integer used in  $\Delta$ . Let  $n_0$  be this greatest occurring integer. Define  $N = \{0, 1, \dots, n_0, n_0 + 1\}$ .

We define the WCMA  $\mathcal{A}' = \langle Q \times N, \Sigma, \Delta', (q_0, 0), F \times N \rangle$  where  $\Delta' = \bigcup_{\delta \in \Delta} S_\delta$  where, for  $\delta = (q, a, c, \uparrow^+, m, q')$ :

$$\begin{aligned} S_\delta = & \{(q, i), a, \perp, (q', m) \mid i \in N, 0 \models c\} \\ & \cup \{(q, i), a, (q'', l), (q', j) \mid i, l \in N, l \models c, j = \min(l + m, n_0 + 1)\} \end{aligned}$$

and for  $\delta = (q, a, c, \downarrow, m, q')$ :

$$\begin{aligned} S_\delta = & \{(q, i), a, \perp, (q', m) \mid i \in N, 0 \models c\} \\ & \cup \{(q, i), a, (q'', l), (q', m) \mid i, l \in N, l \models c\} \end{aligned}$$

Again, showing that these automata recognise the same language is a straightforward induction. Note that this construction is in PTIME since it is assumed that integers in CCAs are given in unary.

It follows immediately from the corresponding result for CCA that the emptiness problem for WCMA is EXPSpace-complete. We note that although here we have provided an equivalence with CCA, this is also straightforward to show by direct reduction to weak multicounter automata: the counters can be used to count how many data values were last seen in each state.<sup>2</sup>

<sup>2</sup> Indeed, our reasons for calling these types of CMA “weak” stems from the equivalence between these weak CMA and weak multicounter automata, mirroring the equivalence between (strong) CMA and (strong) multicounter automata

## B Proof that NDCMA reduce to a WSTS

We assume familiarity with the theory of well-structured transition systems, as described in [6].

For this we will fix a nested data set,  $\langle \mathcal{D}, \text{pred} \rangle$  of level  $k$ , and we make the following observation: a class memory function  $f : \mathcal{D} \rightarrow Q_\perp$  can be interpreted as a labelling of the forest  $\mathcal{D}$  with labels from  $Q_\perp$ . To simplify the following discussion, we add a distinguished “level-0” data value, which is the parent of all of the level-1 data values; this means we are working only over a tree, rather than a forest. Further, we note that by the definition of the transition function, if  $f$  is a class memory function obtained as part of a reachable configuration  $(q, f)$ , then  $f(d) = p \in Q$  implies  $f(\text{pred}(d)) \neq \perp$  (here we assume that the class memory function has some distinguished symbol for the level-0 data value). Hence, in our labelled tree  $\mathcal{D}_f$ , whenever we reach a node labelled  $\perp$ , all of that node’s descendants are also labelled  $\perp$ . Thus, each reachable class memory function gives rise to an unordered labelled finite tree,  $T_f$  (with labels from  $Q$  plus a distinguished symbol for the root) of depth  $\leq k + 1$ . Further, given two class memory functions  $f$  and  $f'$  such that  $T_f = T_{f'}$ , it is clear that  $f$  is equivalent to  $f'$  up to automorphism of  $\mathcal{D}$  (i.e. renaming of data values).

Given an NDCMA  $\mathcal{A} = \langle Q, \Sigma, \Delta, q_0, F \rangle$  we construct the WSTS  $\mathcal{S} = \langle S, \rightarrow, \leq \rangle$  as follows:

- $S = Q \times \Phi$  where  $\Phi$  is the set of unordered labelled finite trees of depth  $\leq k + 1$  (with labels from  $Q$  plus a distinguished symbol for the root)
- $\rightarrow$  is defined as follows:  $(q, T) \rightarrow (q', T')$  iff there are Class Memory Functions  $f_T$  and  $f_{T'}$  s.t.  $T_{f_T} = T$  and  $T_{f_{T'}} = T'$  and  $(q', f_{T'})$  is a valid successor configuration (in  $\mathcal{A}$ ) of  $(q, f_T)$ .
- $\leq$  is defined as follows:  $(q, T) \leq (q', T')$  iff  $q = q'$  and there is a tree homomorphism from  $T$  to  $T'$ .<sup>3</sup>

That  $\rightarrow$  is finite-branching is obvious. That  $\leq$  is a qo is immediate from composition of homomorphisms. That it is a well-quasi-order is given by Lemma 7 in [12]. The upward-compatibility condition is straightforward: if  $(p, R) \leq (q, T)$  and  $(p, R) \rightarrow (p', R')$  then  $p = q$  and there is some  $\delta \in \Delta$  such that applying  $\delta$  from configuration  $(p, f_R)$  yields configuration  $(p, f_{R'})$ . Applying the same  $\delta$  from configuration  $(q, T)$  will yield an appropriate  $(p', T') \geq (p', R')$ .

Decidability of  $\leq$  is also straightforward: to decide whether  $(q, T) \leq (q', T')$  first check  $q = q'$ , then simply check each possible injection from  $T$  to  $T'$  for being a homomorphism.

The only remaining property to check is that we have an effective pred-basis. We do this by showing an effective pred-basis wrt each transition in  $\mathcal{A}$ . The union of these can then be taken. Formally, we compute  $pb(s)$  as  $pb(s) = \bigcup_{\delta \in \Delta} S_{s, \delta}$ .

<sup>3</sup> A tree homomorphism from  $T_1$  to  $T_2$  is a function,  $\phi$ , mapping each node of  $T_1$  to a node of  $T_2$  which maps the root node to the root node, and preserves *pred* and labels. (i.e.  $\phi(\text{pred}(n)) = \text{pred}(\phi(n))$  and  $\text{lab}(n) = \text{lab}(\phi(n))$ .)

We now define these  $S_{s,\delta}$ . Assume  $s = (q, T)$ . If  $\delta$  is a transition which does not go to state  $q$ ,  $S_{s,\delta} = \emptyset$ . Otherwise,  $\delta = p \xrightarrow{a, (i, \bar{s})} q$  for appropriate  $p$ ,  $a$ ,  $i$ , and  $\bar{s}$ . In this case  $S_{s,\delta} = \{T_{\rho, \bar{s}} : \rho \in R_{T, \bar{s}}\}$ , where

$R_{T, \bar{s}}$  = the set of downward-paths,  $n_0, n_1, \dots, n_j$ , from the root, such that the labels of this path (excluding the root's label) give some non-empty prefix of  $q^k$  and if  $s_k = \perp$  then  $n_k$  has no children other than (possibly)  $n_{k+1}$

$T_{\rho, \bar{s}}$  is then defined as follows:

- If  $\rho$  is a path of  $i$  nodes (i.e. the path labels were  $q^i$ ), then  $T_{\rho, \bar{s}}$  is constructed by replacing the labels of the nodes in  $\rho$  with the corresponding element of  $\bar{s}$ . If such an element is  $\perp$  then the node is deleted (note that in this case by choice of  $\rho$  any child nodes must also be in  $\rho$  and labelled with  $\perp$ , so are also deleted)<sup>4</sup>
- If  $\rho$  is a path of  $1 \leq j < i$  nodes, then  $T_{\rho, \bar{s}}$  is constructed by replacing the labels of the nodes in  $\rho$  with the corresponding element of  $\bar{s}$ , and adding nodes labelled with the remaining elements of  $\bar{s}$  as a branch off the last node in  $\rho$ . Any nodes labelled with  $\perp$  are then deleted (again, due to the choice of  $\rho$  and constraints on transitions this must result in obtaining a tree).

## C Proof that HOMCA are equivalent to HOMCA'

We note that HOMCA and HOMCA' clearly have equal expressivity in the level-1 and level-2 cases.

In general, to simulate a HOMCA' with a HOMCA is straightforward: the only extra work that needs be done is prevent  $store_i$  or  $new_i$  operations happening when they are not allowed to in a HOMCA'. This can be done by storing which of the current configuration's multisets are currently undefined in the states.

Simulating a HOMCA with a HOMCA' requires dealing with configurations  $(q, m_1, \dots, m_k)$  where for some  $i < j < k$   $m_j = \perp$  and  $m_i \neq \perp \neq m_k$ . In particular,  $m_1, \dots, m_{j-1}$  may then be split off into a new level- $j$  multiset. We describe how this can be dealt with in the level-3 case, and extensions of this method can be used to deal with higher levels. The main idea will be, when in configuration  $(q, m_1, m_2, m_3)$ ,  $m_1 \neq \perp$  and a  $store_2$  operation would occur, to create a new “ghost” level-2 multiset, to copy  $m_1$  across to “under” this multiset. This is done by the HOMCA' repeatedly performing the following operation: removing one letter from  $m_1$ , then storing  $m_1$  in  $m_2$ , storing  $m_2$  in  $m_3$ , loading the “ghost” level-2 multiset and then the loading the copy of  $m_1$ , and adding the letter removed from  $m_1$  to the copy of  $m_1$ . At the end of this process, the original  $m_1$  is marked as “inactive”, and will not be used again. If anything was

<sup>4</sup> Note that the union of the  $T_{\rho, \bar{s}}$  for these paths actually give  $Pred((q, T))$ . The upward closure is obtained from the shorter paths in the next case.



left in  $m_1$ , it will still be there at the end of the run, so the run will not be an accepting one. When, later, a new level-2 multiset would be created, this can be done by simply removing the “ghost”-marking from the level-2 multiset already created.

Formally, we add to the multiset alphabet,  $\Gamma$ , new letters of the form  $(active, i)$ ,  $(inactive, i)$ ,  $(mf, i)$ ,  $(mt, i)$ ,  $(current, i)$ , and  $(ghost, i)$ , where  $i \in \{1, 2, 3\}$ . The  $mf$  and  $mt$  tags will be used to keep track of which multisets we are “moving from” and “moving to” in the folding-and-unfolding process described above. Whenever a level- $i$  multiset is made, it is immediately populated (using a series of  $\epsilon$ -transitions) to contain the level- $(i - 1)$  multiset  $\{(active, i)\}^{i-1}$ .<sup>5</sup> We can then, when a level- $i$  multiset is loaded, check for the “active” tag by making the counter operations:

$$load_{i-1} \cdot \dots \cdot load_1 \cdot dec_{(active, i)} \cdot inc_{(active, i)} \cdot store_1 \cdot \dots \cdot store_{i-1}$$

We abbreviate this sequence of operations to  $check_{active, i}$ . Now, when a  $store_2$  operation would happen with a non- $\perp$  level-1 multiset open, we make the following sequence of counter operations:

- First, the level-1 multiset is changed from containing  $(active, 1)$  to containing  $(mf, 1)$ , and stored in the level-2 multiset (with operations  $dec_{(active, 1)} \cdot inc_{(mf, 1)} \cdot store_1$ )
- The level-2 multiset is also marked with operations  $load_1 \cdot dec_{(active, 2)} \cdot inc_{(mf, 2)} \cdot store_1$
- The “ghost” level-2 multiset is then initialised with operations  $store_2 \cdot new_2 \cdot new_1 \cdot inc_{(mt, 2)} \cdot store_1$
- The new copy of  $m_1$  is initialised with operations  $new_1 \cdot inc_{(mt, 1)}$
- Then we return to the original copy of  $m_1$  with operations  $store_1 \cdot store_2 \cdot load_2 \cdot check_{mf, 2} \cdot load_1 \cdot check_{mf, 1}$
- We then non-deterministically run the following operations any number of times (with possibly different  $\gamma \in \Gamma$  on each iteration):

$$dec_\gamma \cdot store_1 \cdot store_2 \cdot load_2 \cdot check_{mt, 2} \cdot load_1 \cdot check_{mt, 1} \cdot inc_\gamma \cdot store_1 \cdot store_2 \cdot load_2 \cdot check_{mf, 2} \cdot load_1 \cdot check_{mf, 1}$$

- After this has run non-deterministically many times, we assume that all of  $m_1$  has been copied, and change its marking to inactive, with the operations  $dec_{(mf, 1)} \cdot inc_{(inactive, 1)} \cdot store_1$
- We then return the marking of  $m_2$  to normal and store it as required, with the operations  $load_1 \cdot dec_{(mf, 2)} \cdot inc_{(active, 2)} \cdot store_1 \cdot store_2$
- Finally, we correct the markings of the new multisets with the following operations:  $load_2 \cdot load_1 \cdot dec_{(mt, 2)} \cdot inc_{(ghost, 2)} \cdot store_1 \cdot load_1 \cdot dec_{(mt, 1)} \cdot inc_{(active, 1)}$

When a  $new_2$  operation would subsequently happen, the current level-1 multiset at the time would have its marking changed from  $(active, 1)$  to  $(current, 1)$ . This

<sup>5</sup> Following notation from [3], we write  $\{a\}^j$  to mean the level- $j$  multiset  $\{\dots \{a\} \dots\}$ .

is then stored, and the level-2 multiset has its marking changed from  $(ghost, 2)$  to  $(active, 2)$ . Finally, a  $load_1$  operation is made, and the loaded multiset has its marking changed from  $(current, 1)$  to  $(active, 1)$ . At the end of the run, the automaton is able to remove any *active*, *inactive*, *current*, and *ghost* markings.

This construction guarantees that every element is moved across correctly, because if not the ignored element will still be present at the end of the run, violating hereditary emptiness. In the weak case, this cannot be enforced, and this operation may behave “lossily”. However, this is not a problem: if it is possible for there to be a “lossy” run, then it is possible for there to be a non-lossy run. If there is a lossy run which is accepting, then the corresponding non-lossy run must also be accepting, since increments of counters cannot prevent acceptance.

We have just given a basic example, but the same ideas can be used for the constructions necessary for dealing with  $store_i$  operations with  $i > 2$ . In this example we only needed to move across a single level-1 multiset, which we did by iterating over each of the elements in the multiset. When moving a level-2 multiset, each level-1 multiset must be moved, by iterating over them, using the method described here for each one.

## D Proof that NDCMA are equivalent to HOMCA

For this proof we make use of some slight syntactic sugar for NDCMA, which does not change their power or other properties. Instead of the transition function just giving a single state which the automaton moves to and which class memory function stores for the current data value and all of its ancestors, the transition function can specify the class memory function’s newly stored value for the input data value and each of its ancestors individually. We also add a level-0 data value that is the parent of every level-1 data value (and we require that only data values of level-1 and above can actually be read). Formally, the transition function is now  $\delta = \bigcup_{1 \leq i \leq l} \delta_i$  where each  $\delta_i$  is a function:

$$\delta_i : Q \times \Sigma \times (\{i\} \times (Q_{\perp})^{i+1}) \rightarrow \mathcal{P}(Q \times Q^{i+1})$$

When the automaton is in configuration  $(q, f)$ , reads input letter  $\begin{pmatrix} a \\ d \end{pmatrix}$ , and takes transition  $(q, a, (i, \begin{pmatrix} s_0 \\ \vdots \\ s_i \end{pmatrix})) \rightarrow (q', \begin{pmatrix} t_0 \\ \vdots \\ t_i \end{pmatrix})$ , the automaton moves to configuration  $(q', f')$  where  $f' = f[pred^i(d) \mapsto t_0, pred^{i-1}(d) \mapsto t_1, \dots, d \mapsto t_i]$ . Notationally, we may write this transition as:

$$p \xrightarrow{a, (i, \begin{pmatrix} s_0 \\ \vdots \\ s_i \end{pmatrix})} q, \begin{pmatrix} t_0 \\ \vdots \\ t_i \end{pmatrix}$$

A level- $k$  NDCMA  $\langle Q, \Sigma, \delta, q_0, F_L, F_G \rangle$  with this syntactic sugar can be simulated by a level- $k$  NDCMA with state-set  $Q^{k+1}$ , by using the  $(j+1)^{th}$  component

of the state to store the specified location for the level- $j$  ancestor of the read data value.

We prove the result by showing the two simulation separately.

**Proposition 20.** *Given a strong (resp. weak) NDCMA  $\mathcal{A}$ , a strong (resp. weak) HOMCA'  $\mathcal{A}'$  of the same level can be constructed such that  $\mathcal{L}(\mathcal{A}') = \text{str}(\mathcal{L}(\mathcal{A}))$*

*Proof.* Following notation from [3], for a letter of the multiset alphabet  $a$  we write  $\{a\}^1$  for a level-1 multiset containing just the letter  $a$  (once), and  $\{a\}^{n+1}$  for a level- $(n+1)$  multiset containing just the multiset  $\{a\}^n$  (once). Data values are then represented as follows:

- level- $k$  data values are stored as letters in level-1 multisets: a letter  $(s, k)$  represents a level- $k$  data value remembered as being in state  $s$ .
- level- $(k-1)$  data values are stored as level-1 multisets, containing a letter for each nested data value, and a letter  $(s, k-1)$  (where  $s$  is the state the data value is remembered as being in).
- level- $(n-1)$  data values are stored as level- $(k-n+1)$  multisets, containing a level- $(k-n+1)$  multiset for each nested data value, and a multiset  $\{(s, n-1)\}^{k-n}$ .

Let  $\mathcal{A} = (Q, \Sigma, \Delta, q_0, F_L, F_G)$  be a level- $k$  NDCMA, we define the HOCMA  $\mathcal{A}'$  as follows:

- $\mathcal{A}'$  is a level- $k$  HOCMA
- The multiset alphabet is  $Q \times \{0, \dots, k\}$
- The states are  $Q \uplus \{q_F\} \uplus \bigcup_{\delta \in \Delta} T_\delta$  (where  $T_\delta$  is some set of states which will be used in simulating  $\delta$ -transitions).
- The initial state is  $q_0$
- $q_F$  is the only accepting state
- The transitions are given as follows:

- For each transition  $\delta = p \xrightarrow{a, (i, \begin{pmatrix} p_0 \\ \vdots \\ p_i \end{pmatrix})} q, \begin{pmatrix} q_0 \\ \vdots \\ q_i \end{pmatrix}$  we have a sequence of transitions starting from  $p$  and ending in  $q$ , going through the states  $T_\delta$ , which make counter operations as follows:

- \* We assume that at the start of these transitions, the configuration has multisets  $(\perp, \dots, \perp, m_k)$ . (i.e. only the top-level multiset is defined). (If  $p$  is the initial state this may not be the case, but the way to amend the following construction when the 0th level data value hasn't been seen is straightforward).
- \* We make counter operations:

$$\text{load}_{k-1} \cdot \text{load}_{k-2} \cdots \text{load}_1 \cdot \text{dec}_{(p_0, 0)} \cdot \text{inc}_{(q_0, 0)} \cdot \text{store}_1 \cdot \text{store}_2 \cdots \text{store}_{k-1}$$

to ensure that the level-0 data value is in the correct place, and to update it.

- \* We then  $load_{k-1}$  to select a level-1 data value which will be used for this transition, unless  $p_1 = \perp$  in which case  $new_{k-1}$  is run.
- \* We then do operations

$$load_{k-2} \cdots load_1 \cdot dec_{(p_1,1)} \cdot inc_{(q_1,1)} \cdot store_1 \cdot store_2 \cdots store_{k-2}$$

to correctly check and update the level-1 data value. (Unless  $p_1 = \perp$ , in which case the  $load_i$  operations are replaced with  $new_i$ , and the  $dec$  operation is omitted.)

- \* This process is repeated until we have checked the level- $i$  data value by doing  $dec_{(p_i,i)} inc_{(q_i,i)}$ . Then all of the opened multisets are stored, up to  $store_{k-1}$ .

All but the first of these transitions are labelled with  $\epsilon$  (and the first is labelled with  $a$ )

- From states in  $F_G$  we have an  $\epsilon$  transition to the state  $q_F$
- From state  $q_F$  there are transitions to make any load or store operation, and to make any  $dec_{(s,i)}$  where  $s \in F_L$ .

It is a straightforward induction to show that  $\mathcal{A}'$  simulates  $\mathcal{A}$ . If  $\mathcal{A}$  is weak, then it is clear that from  $q_F$  any configuration can be reduced to a hereditarily empty one, hence a weak HOCMA' is enough.

**Proposition 21.** *Given a strong (resp. weak) HOCMA'  $\mathcal{A}'$ , a strong (resp. weak) NDCMA  $\mathcal{A}$  of the same level can be constructed such that  $str(\mathcal{L}(\mathcal{A})) = \mathcal{L}(\mathcal{A}')$*

*Proof.* We note that by definition of HOCMA', the only reachable configurations are of the form  $(q, m_1, \dots, m_n)$  such that for some  $i \leq n$  we have that for all  $j < i$ ,  $m_j = \perp$ , and for all  $j' \geq i$ ,  $m_{j'} \neq \perp$ . (Hence the first counter operation used must be  $new_n$ .) In the following construction we will use the remembered location of the level-0 data value to track the value of this  $i$ .

Given a level- $k$  HOMCA  $\mathcal{A}' = \langle Q', \Sigma, \Delta', q'_0, F' \rangle$  over a multiset alphabet  $A$ , we construct a level- $k$  NDCMA  $\mathcal{A} = \langle Q, \Sigma, \Delta, q_0, F_L, F_G \rangle$  that recognises the same string-projection as  $\mathcal{A}'$  as follows:

- $Q = Q' \uplus A \uplus \{(1), \dots, (k)\} \uplus \{\bullet, \circ\} \uplus \{q_{dead}\}$ . Here the states  $(i)$  will be used to keep track of the smallest  $i$  such that  $m_i \neq \perp$  in the current configuration, and the state  $\bullet$  will be used to keep track of which data values are representing multisets currently “active” (i.e. equal to one of the  $m_i$ ), while the state  $\circ$  will store the “inactive” multisets.

$$- q_0 = q'_0$$

$$- F_G = F'$$

$$- F_L = \{q_{dead}\} \cup \{(1), \dots, (k)\} \cup \{\bullet, \circ\}$$

- If  $p \xrightarrow{a, op} q \in \Delta'$  then we have a transition  $p \xrightarrow{a, (i, \begin{pmatrix} p_0 \\ \vdots \\ p_i \end{pmatrix})} q, \begin{pmatrix} q_0 \\ \vdots \\ q_i \end{pmatrix}$  where  $i$ ,

$\begin{pmatrix} p_0 \\ \vdots \\ p_i \end{pmatrix}$  and  $\begin{pmatrix} q_0 \\ \vdots \\ q_i \end{pmatrix}$  are such that:

- If  $op = new_k$  then  $i = 0$  and  $p_0 = \perp$  and  $q_0 = (k)$
- If  $op = new_m$  when  $m \neq k$  then  $i = k - m$  and:
  - \*  $p_0 = (m + 1)$  and  $q_0 = (m)$
  - \* for  $0 < j < i$   $p_j = \bullet = q_j$
  - \*  $p_i = \perp$  and  $q_i = \bullet$
- If  $op = inc_b$  ( $b \in A$ ) then  $i = k$  and
  - \*  $p_0 = (1) = q_0$
  - \* for  $0 < j < k$   $p_j = \bullet = q_j$
  - \*  $p_k = \perp$  and  $q_k = b$
- If  $op = dec_b$  ( $b \in A$ ) then  $i = k$  and
  - \*  $p_0 = (1) = q_0$
  - \* for  $0 < j < k$   $p_j = \bullet = q_j$
  - \*  $p_k = b$  and  $q_k = q_{dead}$
- If  $op = load_m$  ( $m < k$ ) then  $i = k - m$  and:
  - \*  $p_0 = (m + 1)$  and  $q_0 = (m)$
  - \* for  $0 < j < i$   $p_j = \bullet = q_j$
  - \*  $p_i = \circ$  and  $q_i = \bullet$
- If  $op = store_m$  ( $m < k$ ) then  $i = k - m$  and:
  - \*  $p_0 = (m)$  and  $q_0 = (m + 1)$
  - \* for  $0 < j < i$   $p_j = \bullet = q_j$
  - \*  $p_i = \bullet$  and  $q_i = \circ$

It is again straightforward to show this simulates the run of  $\mathcal{A}'$ . The acceptance condition is checked by  $F_L$  not including  $A$ : the only way for the multiset to not be hereditarily empty is for an increment never to have a corresponding decrement, and this (and only this) can leave a data value in  $A$ . If we have a weak HOCMA', then we do not need to check this and can let  $F_L$  be the whole set of states.

This completes the proof of theorem 17. We note that our earlier theorem 9 is now a corollary of emptiness of HOMCA being undecidable (shown in [3]).